Feature

Folding research recruits unconventional help

A denatured protein chain can find its well-ordered three-dimensional structure, the native state, in under a second, using only the information contained in the sequence. For researchers, however, the prediction of structures from sequences is a hard problem, so they are now recruiting all the help they can get, including idle computers and game consoles, game players, and little hints from evolution. **Michael Gross** reports.

Protein folding is one of the miracles of nature that human technology finds quite difficult to follow. Ever since the classic ribonuclease A experiments of Christian Anfinsen in the 1960s it has been clear that the amino acid sequence of a polypeptide chain determines the unique three-dimensional folded conformation it will adopt under physiological conditions. Even though we now know that some proteins remain intrinsically disordered, and some may 'fold around' a ligand, it is still true that a protein's structure, and hence its function, is somehow encoded in the sequence of the amino acids. As the theoretician Cyrus Levinthal pointed out early on, there is an astronomical number of wrong conformations which the chain cannot possibly try out in a reasonable time, so there must be mechanisms that allow polypeptide chains to find the native state encoded in their sequence. Folding researchers have elucidated some of these mechanisms in the last decades, but so far haven't been able to decipher the code and therefore aren't generally able to read a sequence and predict what shape it will adopt.

An early approach to the problem was to build bigger computers dedicated to simulations of folding, a development that culminated in the development of IBM's Blue Gene, first announced in 1999 with the explicit target of tackling protein folding, but it didn't crack the prediction problem once and for all. By the turn of the millennium, computer simulations of the protein movements could only just cover a microsecond, while it was known from experimental studies that the relevant folding reactions happened on the millisecond timescale.

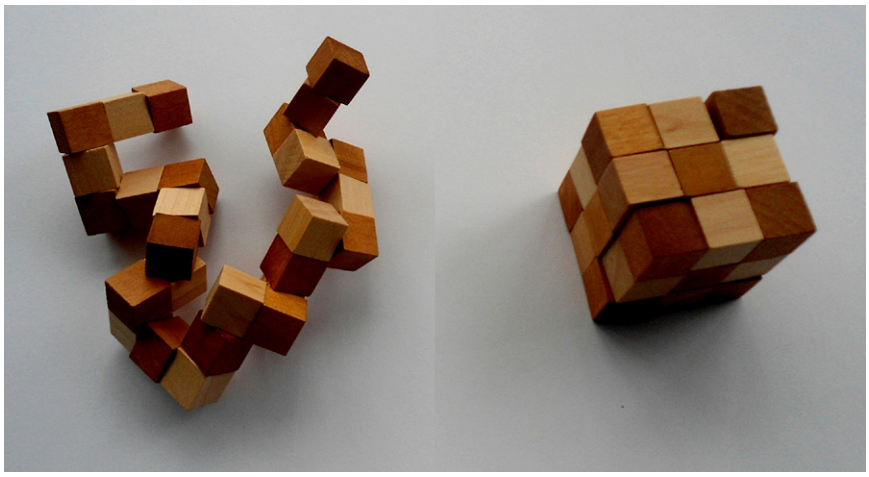
Folding at home

Based on that shortfall of computing power and the increasing availability of PCs connected via the internet, the group of Vijay Pande at Stanford University developed a distributed computing program called Folding@home, using the now widely adopted practice of chopping a problem into small parcels and farming them out to many computers that are online but idle (as, for instance, most computers in universities are, most of the time).

Since 2006, the lab also offers a version of the program that runs on games consoles, which enabled a dramatic increase in the processing power accessible to the program. In September 2007, the program achieved a consistent level above one petaFLOPS (10^{15} floating-point operations per second), as the first computing system of any kind to do so (the fastest supercomputer at the time was Blue Gene with just over a quarter of that power). In November 2011, Folding@home passed the milestone of six petaFLOPS.

The simulation of protein dynamics by the Folding@home software is based on the observation that protein chains populate certain free energy minima for a period of time, and then quickly move on to another minimum. Pande's group uses so-called Markov state models to find connections between these minima and map the likelihood of transitions. In 2010, the group used this approach combined with the distributed computing power of Folding@home to simulate the folding of the 39-residue protein NTL9, which takes about 1.5 milliseconds (J. Am. Chem. Soc. (2010), *132*, 1526–1528). The time span of this simulation was a thousand times longer than that achieved by other methods.

Pande's group uses this approach to address a range of biomedically relevant issues around protein folding, including diseases that involve misfolded proteins (such as Alzheimer's, Parkinson's and Huntington's disease), the fundamental questions of protein folding mechanisms, and the prediction of unknown structures from sequences, which feed into commercial drug design. "Through the Folding@home distributed computing project, we have been able to muster an unparalleled computational resource for studying protein folding,



Folding puzzle: The rearrangement of a linear polymer into a compact three-dimensional shape proceeds autonomously in nature but is still puzzling biochemists. (Photo: Michael Gross.)

Folding fun: Tens of thousands of online game players around the world have contributed to folding research via the game Foldit, developed at the University of Washington at Seattle. (Photo: Mohini Patel Glanz.)

allowing us to study complex systems on timescales thousands of times longer than would otherwise be possible," says Pande. The group makes all datasets generated from the research available on request.

Folding game

Meanwhile, the group of David Baker at the University of Washington at Seattle had developed an algorithm called Rosetta for the prediction of small protein structures, and also set up distributed computing (Rosetta@home) to provide additional computing power for the prediction work. Participants who have this program installed on their computers can watch how the protein chain gradually finds its native conformation. It so happened that some participants watched the process and spotted possible arrangements that would improve the energy-efficient packing, but weren't able to interact with the program, finding themselves in the situation of someone watching a game show and shouting at their TV set.

From this frustration, the idea was born to create an interactive version of the prediction program, where human intuition can help to wriggle parts of the protein into the right places. Together with computer scientist David Salesin and games developer Zoran Popovic (both also at the University of Washington), Baker's group developed a multiplayer online game called Foldit, releasing the first public beta version in May 2008.

Participation in the game doesn't require any knowledge of biochemistry. As in commercial computer games, players can rise through various levels of difficulty depending on their success, and the first few levels are designed to teach them the skills required to twist and turn polypeptide chains into energetically optimised shapes. When they reach the level where they can tackle real scientific problems, players can choose to play on their own or in a team, competing with other players and teams. Success is rewarded with a points system, and there are also social components such as chat and wiki features, where the players can swap notes and get to know each other.

Two years after the launch, and after 57,000 players had participated, the Baker group reported the first scientific results emerging from the game (Nature (2010), *466*, 756–760). They could show that the highest-ranked players were better at predicting unknown structures than the best algorithms available. Their success can be readily measured using protein structures that are about to be solved experimentally, or have been solved but aren't published yet. Overlaying experimental and predicted structures, one can calculate differences that can be summarised as a mean deviation, which is a single figure providing a measure of prediction quality.

Which human qualities enable the most proficient players to beat the computer algorithms? One significant problem of the existing programs is that, while they are very efficient at sampling readily accessible improvements, they may fail when they get trapped in a local minimum, i.e. a conformation that is better than everything around it and would require major rearrangement to access the real structure. A human player can decide to rip up parts of the existing structure and piece it together differently, using visual intuition to find alternatives that may be very remote for a computer program that would have to find them in a stepwise fashion.

Moreover, the involvement of thousands of players is bound to generate many different procedures and strategies, so the game community as a whole will have a wider range of options than a computer algorithm may have. Not all of these will lead to useful solutions, but the diversity improves the chances that someone will find the right fold eventually.

And where does the human player struggle to compete with the algorithms? The most difficult step appears to be the 'blank slate' when the player has to start from a fully extended chain and anything might be possible. Thus, there is a case for combining automatic procedures with manual manipulations to get the best of both worlds for folding research.

Folding recipes

Following the first prediction successes of Foldit players, Baker's group decided to enable the players to record their methods in the form of recipes, which they can collect and share (or keep to themselves) in their personal cookbooks. The researchers found that the sharing and modification of recipes between the players led to a rapid evolutionary improvement of methods, which became an interesting subject of study in its own right (Proc. Natl. Acad. Sci. USA (2011), *108*, 18949–18953).

The most widely used and most often copied and modified recipe of the Foldit players is called Blue Fuse. Essentially it involves iterative rounds

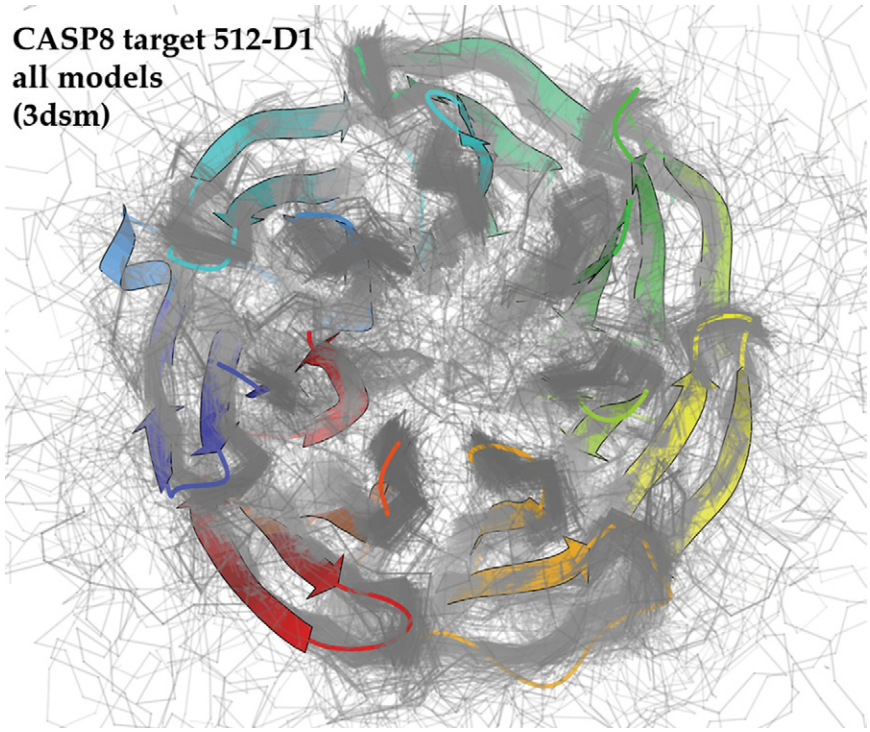
of compressing and relaxing the structure. It turned out that Blue Fuse operates in a similar way to a new algorithm that researchers in Baker's lab were developing at the same time. Comparing the performance of the player recipe to the scientists' algorithm, the researchers found that, within the limitations of the game, Blue Fuse is more efficient, while in the laboratory setting the computer algorithm outperforms it.

Baker's group and the Foldit players could celebrate a major scientific success recently, when the game helped to crack the crystal structure of a protein (the monomeric retroviral protease of the Mason-Pfizer monkey virus) that had eluded structural biologists for more than a decade after crystallographic datasets had been collected. As is often the case in protein crystallography, solving the phase problem — a loss of information inherent in the method, which is normally amended by comparison of related structures that differ in one aspect — turned out to be the biggest challenge. Several attempts to solve the phase problem for this protease structure by established methods have failed, although NMR structures were available as a starting point.

At 114 residues length, the protein turned out to be a suitable puzzle for Foldit players, who could start from the known NMR structure and modify it manually. After a three-week competition among the players, researchers took stock and found that several of the models produced were good enough to enable crystallographers to solve their structure (Nat. Struct. Mol. Biol. (2011), *18*, 1175–1177).

"I am very excited about what Foldit players have done so far, and look forward to working with them to design new proteins with new functions," David Baker enthuses. "We are currently challenging Foldit players to design flu inhibitors and other potential therapeutics — if they succeed it will be a notable milestone in the changing relation between scientists and society."

The current trend towards distributed computing and crowdsourcing, however, doesn't mean that supercomputers are completely out of the race. Recently, the group of David Shaw, a computer scientist and former investment banker who used his winnings to set up a private research institute at New York, reported progress in the simulation of the folding of 12 small structurally diverse protein domains ranging from 10 to 80 amino acid residues, using a purpose-built supercomputer called Anton (Science (2011), *334*, 517–520), which is claimed to achieve similar performance in molecular dynamics calculations as the Folding@home network.



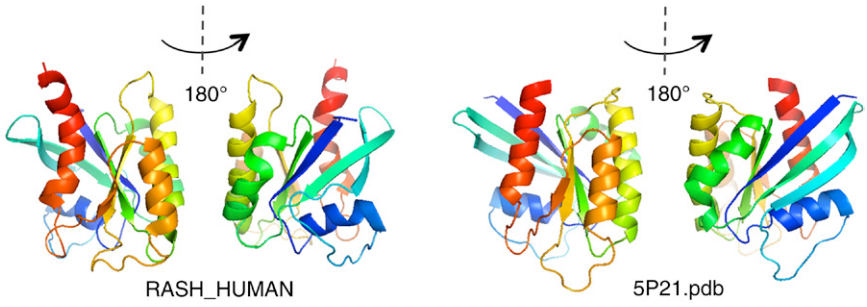
Folding target: The target structure of the CASP8 competition (coloured) compared with the predictions submitted (grey). (Photo: Daniel Keedy, Jane S. Richardson — Source Wikimedia commons.)

Folding Olympics

The trouble with folding simulations under purely mechanistic considerations is that it is hard to compare their performance, as only partial experimental information is available on the 'true' folding mechanism. In folding prediction, however, judging a prediction against a new experimental structure provides a highly objective quality measure.

Based on this idea, the Protein Structure Prediction Center (predictioncenter.org) has held bi-annual competitions of blind prediction known as CASP (Critical Assessment of protein Structure Prediction) since an organising team led by John Moult at the University of Maryland Biotechnology Institute set up the first competition in 1994. CASP10 is due to launch this April. In Addition, the centre has also launched a continuous prediction project, called CASP Roll, in November 2011.

For each CASP competition, the centre uses a recently solved but as yet unpublished protein structure and invites predictions that will then be tested against the experimental result. A full evaluation of the latest round, CASP9, was recently published in a special supplement to the journal Proteins: Structure, Function, and Bioinformatics (vol. *79*, issue S10). Summing up the results of CASP9, Andriy Kryshtafovych, Krzysztof Fidelis, and John Moult conclude that progress since the previous two competitions has been generally modest in comparison to the more rapid progress made in the early years of CASP, when the new competition inspired rapid improvements in methodology. However, they noted a few encouraging trends. Specifically, they observed "some improvement in overall model quality in the midrange of modeling difficulty" and "clear progress in identifying the best model out of five submitted" among other



Prediction progress: Predicted (left) and experimental (right) structure of the Ras G domain studied by Debora Marks *et al.* using only evolutionary constraints for the prediction. (Image reproduced with permission from Marks D.S. *et al.* (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766. doi:10.1371/journal.pone.0028766.)

positive developments. They also note the emergence of new methods that were not yet competing in the 2010 exercise.

Folding evolution

Apart from the sporting ambition of solving a puzzle, another factor motivating prediction efforts is the fact that the number of gene sequences known is still several orders of magnitude larger than the number of high-resolution structures of proteins. Given the recent million-fold increase in cost-efficiency of genome sequencing (Curr. Biol. (2011), *21*, R294–R297), this gap is likely to grow even wider.

One might see the millions of gene sequences and the protein sequences derived from them piling up in databases as a frustrating accumulation of unsolved mysteries. The more optimistic view is to see them as a reservoir of evolutionary information that could help to solve individual structures.

Several structure predictions have already made use of evolutionary constraints in one way or another. Last December, however, the group of Chris Sander at the Memorial Sloan-Kettering Cancer Center in New York together with Harvard researchers Debora Marks and Lucy Colwell presented the first *de novo* structure predictions that were obtained exclusively on the basis of sequence comparisons with related proteins, without using any other constraints or any of the existing structural information concerning the proteins involved (PLoS One 2011, *6*, e28766).

For each protein structure they wanted to predict, the group used sequence alignments of at least 1,000 related proteins. To assess the feasibility of the approach, they used protein families of which at least one member had a known structure to test the prediction against. They derived distance constraints from the co-evolution of pairs of amino acids that tend to undergo compensating changes in evolution. "We also had to get rid of indirect 'chaining' correlations that don't reflect direct interactions, which we did by adapting a trick from statistical physics working with our friends in Torino," Chris Sander explains. Surprisingly, most of the proteins they tried folded up correctly on the computer, in spite of false-positive signals emanating from pairs of amino-acid residues that co-evolve for other reasons, such as allosteric interactions or functional mechanisms.

Although Sander and others had previously tried a similar approach with a smaller number of homologous sequences and failed, this time the researchers found that the evolutionary information was sufficient to 'predict' the well-known structure of the G domain of the Ras protein correctly without other input. The predicted structure of the Ras protein agreed with the known experimental structure within deviations of 3.5 angstroms, and all the secondary structure elements were in the right places. With a length of 161 amino acids, this protein domain is beyond the size range that standard algorithms could predict *de novo* at the moment.

The researchers also tried the approach on several other protein families with a variety of structures of domains or entire proteins, ranging from the small RNA-binding domain to the large enzyme trypsin and even the transmembrane protein rhodopsin. While not all of these predictions were as successful as the one for the Ras G domain, they showed encouraging results overall and confirmed that the prediction is feasible on the basis of evolutionary constraints alone.

"We set out to mine this deluge of evolutionary information to get at new three-dimensional structures of thousands of proteins with enormous savings of experimental effort and open new doors to protein design — and were stunned when we saw the first results", Debora Marks pointed out.

In practice, of course, the predictions may get even better if the evolutionary method is combined with other types of information already available on the protein in question, e.g. experimental distance constraints.

As the authors conclude, the fact that structures can be predicted from simple evolutionary constraints after removal of all indirect or complex interactions "may be as much a starting point for an exploration of our understanding of the evolution of proteins as it is a route to structure prediction."

Researchers have spent decades refolding proteins *in vitro* and are now learning to fold them *in silico*, but there is still the issue of what actually happens *in vivo*, in the presence of other molecules including folding helpers (molecular chaperones). "So far, molecular dynamics simulations of protein folding are still limited to rather small proteins. But given the progress achieved in recent years, larger systems and larger time scales are within reach," comments Johannes Buchner from the Technical University of Munich. "It will be interesting to see whether in the future the function of molecular chaperones can also be implemented. It may well be the case that chaperones such as Hsp90 play an important role in determining the accessible conformational space for a given protein."

Michael Gross is a science writer based at Oxford. He can be contacted via his web page at www.michaelgross.co.uk